A Predictive Analysis: Study of Turtle Games' Customer Behaviour and Product Sales Performance

By

Nur Athira Binte Mohd Adom

LSE Career Accelerator

Course 3: Advanced Analytics for Organisational Impact

Assignment: Predicting Future Outcomes

Contents

1. Background/context of the business

Turtle Games (TG) is a game manufacturer and retailer catering to thousands of customers worldwide. It sells a wide range of books and games for different platforms. TG has provided data to tackle the business objective of improving overall sales performance by utilising customer trends.

We are working with 2 departments in TG to understand:

1. Customer Trends (Marketing Team)
   a. How users accumulate loyalty points
   b. How customers perceive and receive TG's products
   c. Recommendations for marketing direction
2. Sales Performance of products currently on sale (Sales Team)
   a. Compare sales by product and region
   b. Evaluate reliability of data for making sales predictions

2. Customer Behaviour - Python analysis

The marketing team is focused on understanding the customers' purchasing behaviour and their feedback on products.

2.1. Analytical Approach: Data Processing

2.1.1. Importing & Exploring the Data on Jupyter Notebook

The provided dataset (turtle_review.csv) is imported into the Jupyter notebook. Sense-checking is done to ensure the data is useable and relevant:

1. Previewed headers to view the range of data provided
2. Check for missing values (NA)
3. Removed redundant data
   → "language" and "platform" columns were redundant as all rows had the same value
4. Renames columns
   → To make recalling more intuitive: "`remuneration (k£)`"→ "`remuneration`"
   "`spending_score (1-100)`"→ "`spending_score`"

The useful data provided can be categorised as follows:

Customer Demographic Characteristics	Measures of customer interactions with TG
1) Gender 2) Age 3) Remuneration (k£): income per customer per year in pounds, where k=1000 4) Education level	1) spending_score (1-100): reflective of customer's spending nature and behaviour 2) loyalty_points: based on the point value and monetary value of the purchase 3) review/summary

2.1.2. Preparing the workspace

I imported the following libraries and packages to enable analysis and visualisation of data:

```
# Imports
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.metrics import accuracy_score
from scipy.spatial.distance import cdist
```

Figure 2.0.1.2 Libraries and Packages imported to be used in analysis and visualisation

2.2. Analytical Approach: Predictions with Regression

The marketing team is interested in how customer accumulate loyalty points. This will thus be our dependant variable. We will observe how loyalty points change with age, remuneration, and spending scores to determine whether these can be used for prediction of loyalty points.

2.2.1. Scatter Plot + Linear Regression (by OLS method)

Since the different variables will be compared to Loyalty Points in a similar way, I have chosen to create functions to quickly:

1) Obtain regression table
2) Print useful regression values
3) Visualise scatterplot + regression line

We can create a function to generate the scatter and fit the OLS model

```
In [12]:
# Create function for OLS model and to visualise scatter
def simpleLR_OLS(x,y):
    # Check for linearity with Matplotlib - visualise scatter
    plt.scatter(x, y)

    # Create formula and pass through OLS methods.
    f = 'y ~ x'
    test = ols(f, data = reviews_clean).fit()

    # Print the regression table.
    print(test.summary() )

    return test
```

Useful values to be printed in a separate function to keep the jupyter notebook manageable and easier to trace/refer only to specific data

```
In [13]:
# Create function to print useful values

def simpleLR_OLS_useful(test):
    # Extract the estimated parameters.
    print("Parameters: \n", test.params, "\n")

    # Extract the standard errors.
    print("Standard errors: \n", test.bse, "\n")

    # Extract the predicted values.
    print("Predicted values: \n", test.predict(), "\n")
```

Figure 2.2.1a Screenshot of user-defined functions used

a) Loyalty points vs Spending Score

Useful values:

R-squared = 0.452
Adjusted R-squared = 0.452

low R-squared value

p-value = 2.92e-263

coefficient = 33.0617 (positive)

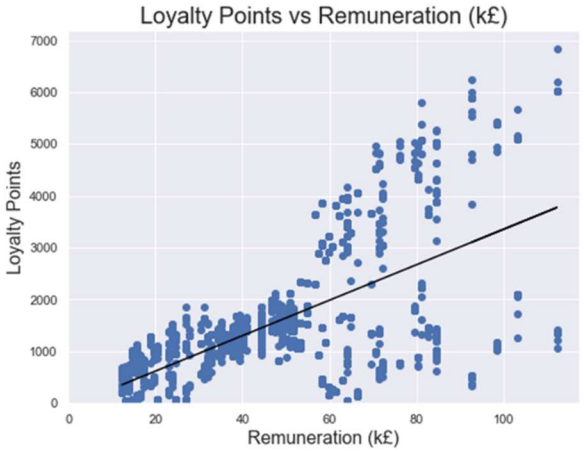
Generally positive relationship between spending and loyalty points

(P>|t|) = 0.000

(significant slope)

95% confidence interval = [31.464 34.659]

b) Loyalty points vs Remuneration



Useful values:

R-squared = 0.380
Adjusted R-squared = 0.379

low R-squared value

p-value = 2.43e-209

coefficient = 34.1878 (positive)

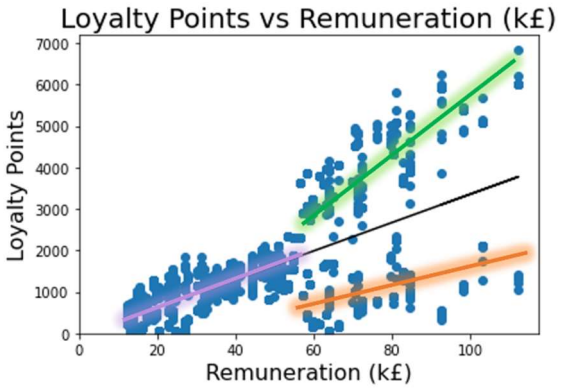
Generally positive relationship between remuneration and loyalty points

(P>|t|) = 0.000

(significant slope)

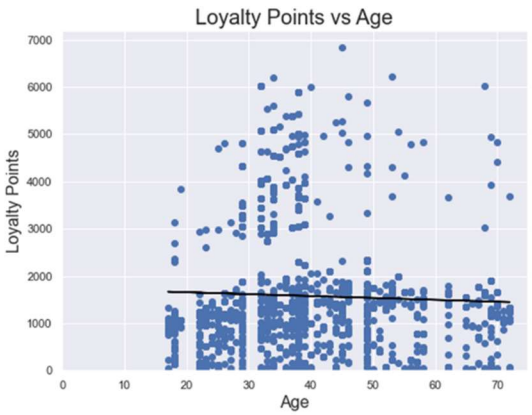
95% confidence interval = [32.270 36.106]

Based on preliminary analysis, we can observe some possible clusters within the sample population that result in varying levels of correlation between remuneration and loyalty points.



The correlation will be reviewed after we evaluate appropriateness of clustering.

c) Loyalty points vs Age



Useful values:

R-squared = 0.002
Adjusted R-squared = 0.001

negligible R-squared value
Model CANNOT be used to predict loyalty points based on customer's age

p-value = 0.0577

coefficient = -4.0128 (negative)

Generally negative relationship between age and loyalty points
However, the **relationship is insignificant**

(P>|t|) = 0.058

(insignificant slope)

95% confidence interval = [-8.157 0.131]

Evaluation

While spending score and remuneration might be used to predict loyalty point accumulation, age is not appropriate for this purpose.

2.2.2. *k*-clustering (by OLS method)

To divide the customer base into specific market segments, we employ *k*-mean clustering to find the optimal cluster number of clusters.

a) Identifying Ideal Number of Clusters

Upon plotting the scatter of Spending Score vs Remuneration, we can expect there to be at least 5 distinct clusters. After applying both the Elbow Method and the Silhouette Method, we identify *k*=5 as the ideal number of clusters.

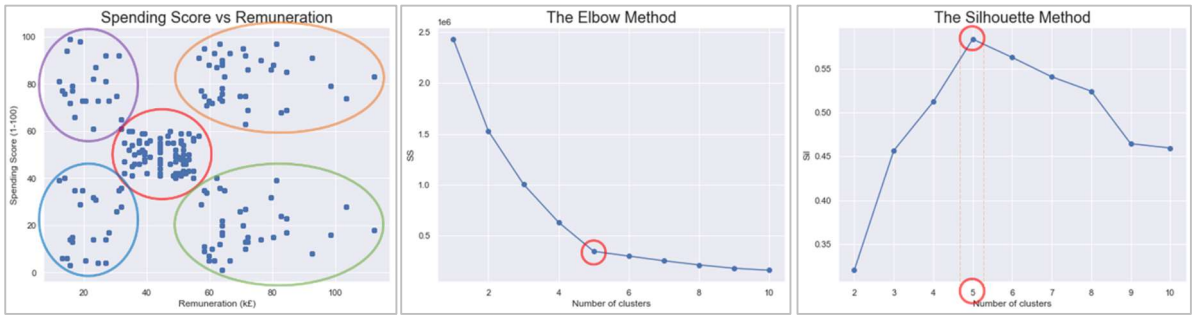


Figure 2.2.2a. Optimal number of clusters identified visually (left), by Elbow Method (middle) and by Silhouette Method (right)

b) Optimising Clustering Model: Evaluate k-means model at different values of k

Besides *k*=5, k-value of 6 and 7 were also tested for effectiveness comparison. The values were chosen from the next highest silhouette scores.

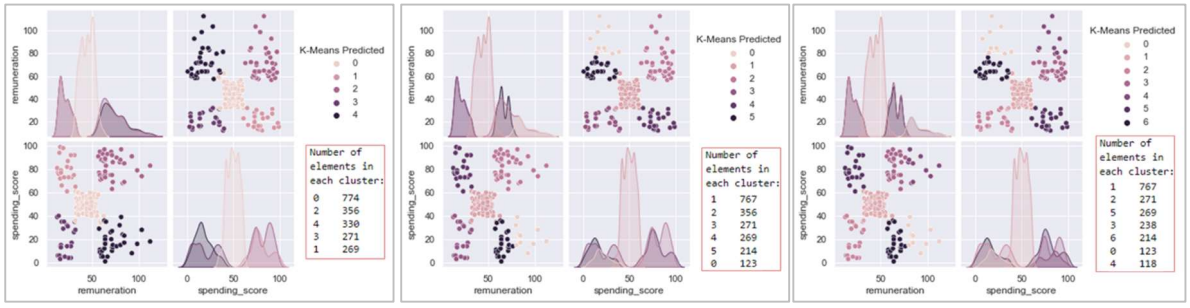


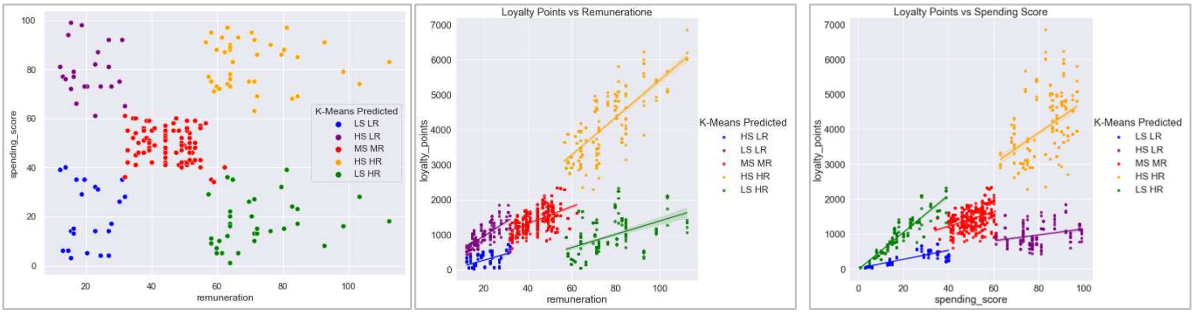
Figure 2.3.2.b Comparison of k-means clustering with k = 5, 6, and 7

Although outliers were more clearly separated when *k*=6 and 7, I decided to use *k*=5 as it would ensure each cluster is a significant market segment (min 13.45%) and reduce complexity of the model while the clusters are still unlabelled.

c) Attempting to Label the Clusters

I also tried to plot the other demographic data to categorise the clusters, but all clusters had a mix of genders, ages, educational level, and products. More information on customers is needed to properly identify clusters.

2.2.3. Observations & Insights



1. Accumulation of Loyalty Points is highly correlated with customer's Spending Score and Remuneration
2. Market can be segmented into 5 key clusters currently grouped by a combination their Remuneration and Spending Score
    a. More investigation will need to be done to understand the clusters better and cater marketing campaigns towards them

2.3. Analytical Approach: Customer Sentiment Analysis by NLP

2.3.1. Importing & Cleaning the Data on Jupyter Notebook

The 'Reviews' and 'Summary' columns are extracted from the same dataset (turtle_review.csv for sentiment analysis. The following steps were done to prepare the data:

1. Imported and loaded necessary libraries and packages
2. Removed duplicates
3. Normalised data:
    a. Used lowercase
    b. Removed punctuation
    c. Keep only English words
    d. Remove stop words
4. Tokenised text

```
import nltk
from wordcloud import WordCloud
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
from nltk.corpus import stopwords, words
from textblob import TextBlob
from scipy.stats import norm
```

Figure 2.3.4. Libraries and Packages imported to be used in analysis and visualisation

2.3.2. Generating Frequency List & Word Cloud

Since the Reviews and Summaries were to be analysed separately but through similar processes, I created a few user-defined functions to reduce repetition. After the words is all reviews have been filtered and tokenised, a frequency list was produced to obtain the 15 most frequent words.

Reviews

	Word	Frequency
1	game	1671
2	great	580
3	fun	552
4	one	530
5	play	502
6	like	414
7	love	323
8	really	319
9	get	319
10	time	291
11	good	289
12	would	280
13	book	273
14	well	269
15	new	264

Summary

	Word	Frequency
1	five	342
2	game	319
3	great	295
4	fun	218
5	love	93
6	good	92
7	four	58
8	like	54
9	expansion	52
10	cute	45
11	book	43
12	one	38
13	awesome	36
14	product	35
15	old	34

2.3.3. Polarity and Subjectivity

Each review and summary was given a polarity and subjectivity rating using the TextBlob library. The reviews and summaries were than sorted to give the 20 most polarising responses

2.3.4. Observations and Insights

1. Customer reviews are generally positive
2. Games and Books are the popular products reviewed
3. There were a handful of false negative reviews (within the top 20)
   a. Unable to fully rely on the automated NLP to correctly evaluate the sentiment in the human language.

Review	Polarity
one of my staff will be using this game soon so i dont know how well it works as yet but after looking at the cards i believe it will be helpful in getting a conversation started regarding anger and what to do to control it	-0.55
i bought this as a christmas gift for my grandson its a sticker book so how can i go wrong with this gift	-0.5
kids i work with like this game	-0.4
my son loves playing this game it was recommended by a counselor at school that works with him	-0.4
this game is a blast	-0.4
i bought this for my son he loves this game	-0.4
was a gift for my son he loves the game	-0.4

Figure 2.3.4. Examples of false negative sentiment polarity assignment

3. Sales Performance – R Analysis

The sales team is focused on understanding the sales performance of different products in different regions. The team would like to understand:

- the impact that each product has on sales
- how reliable the data is (e.g. normal distribution, skewness, or kurtosis)
- what the relationship(s) is/are (if any) between North American, European, and global sales?

3.1. Analytical Approach: Data Processing

3.1.1. Importing & Exploring the Data on RStudio

The provided dataset (turtle_sales.csv) is imported into RStudio. Sense-checking is done to ensure the data is useable and relevant:

1. Previewed dataframe using "as_tibble()" to verify data types and first 10 rows
2. Previewed dataframe using "skim()" to check for missing values
3. Removed redundant data
   - Since the team is only interested in sales by product and region, 'Ranking', 'Year', 'Genre' and 'Publisher' columns were removed
4. Converted Product ID to character for easy processing as categorical data

3.1.2. Exploring data

Categorical Data Variables	Continuous Data Variables
1) Product ID *Although it is numerical, each product is distinct and there is no clear order for the numbering* 2) Platform 3) Genre *(not in use)* 4) Publisher *(not in use)*	1) NA_Sales 2) EU_Sales 3) Global_Sales

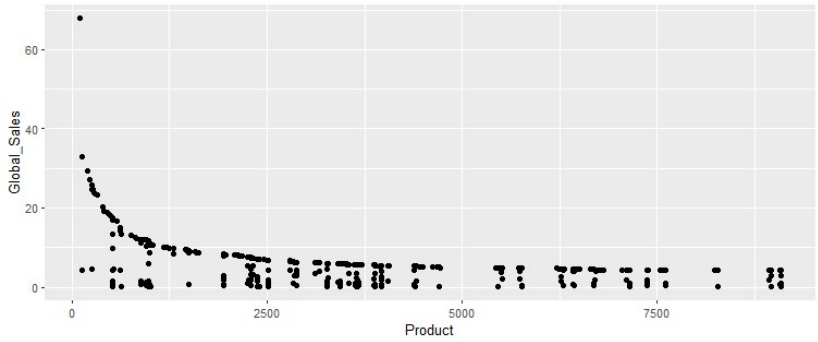
The skim and summary functions are useful in getting an overview of the variables:

- Categorical data: Number of unique values
- Continuous data: Min, Max, Mean, Median, IQR

3.1.3. Visualising data

a) Distribution

a) Global Sales vs Product ID



At first glance, it looks like Global_Sales has an exponential relationship with Product ID towards 0. However, since product ID is categorical, this is not an appropriate visual representation as it is misleading.

b) Global Sales vs NA_Sales

3.4. Final Insights